



UNIONE EUROPEA
Fondo europeo di sviluppo regionale



REPUBBLICA ITALIANA



REGIONE AUTÒNOMA DE SARDIGNA
REGIONE AUTONOMA DELLA SARDEGNA



SARDEGNA
RICERCHE

Next-Gen Bioinformatics

— Moving from wet to dry —

Gianmauro Cuccuru - gmauro@crs4.it

NGS Data Storage requirements



	raw data	pre-processed data
HiSeq 3000	~500 GB	~300 GB
HiSeq 2500	~2,5 TB	~300 GB
HiSeq 2500 rapid	~250 GB	~50 GB

Laboratory Information Management System

The screenshot displays a LIMS interface with a main table of samples and a detailed view of a worksheet for sample RNA17-0001.

SampleID	ClientSID	Type	FASTQ-File	full-analysis	smallRNA-Library	smallRNA-Seq-50SE-10M	Runs	State
RNA17-0060	ITA2	RNA	Success	Success	Success	Success		Received
RNA17-0001			sequencing	pool	ribosome	ribosome		celved
RNA17-0002			sequencing					celved
RNA17-0003	ITA8	RNA	Success	Success	Success	Success		Received
RNA17-0004	ITA12	RNA	Success	Success	Success	Success		Received
RNA17-0005	ITA24	RNA	Success	Success	Success	Success		Received

WorksheetID	Title	Description	Analysis	Analyst	Date	State
WS17-0013			28	sequencing	2017-01-20	Open
WS17-0013			2	sequencing	2017-01-20	Open

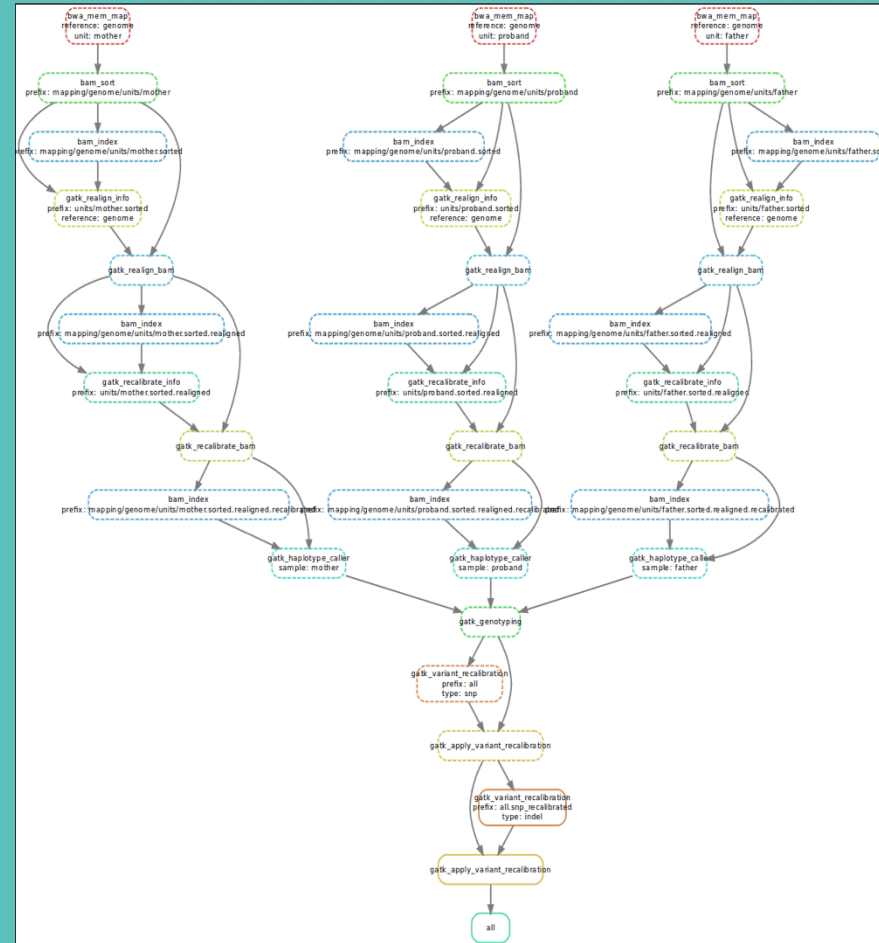
Assign to a worksheet

WORKSHEET Create new Generate FASTQ Pre-Processing data

return

Assign

Downstream analysis



Most tools can:

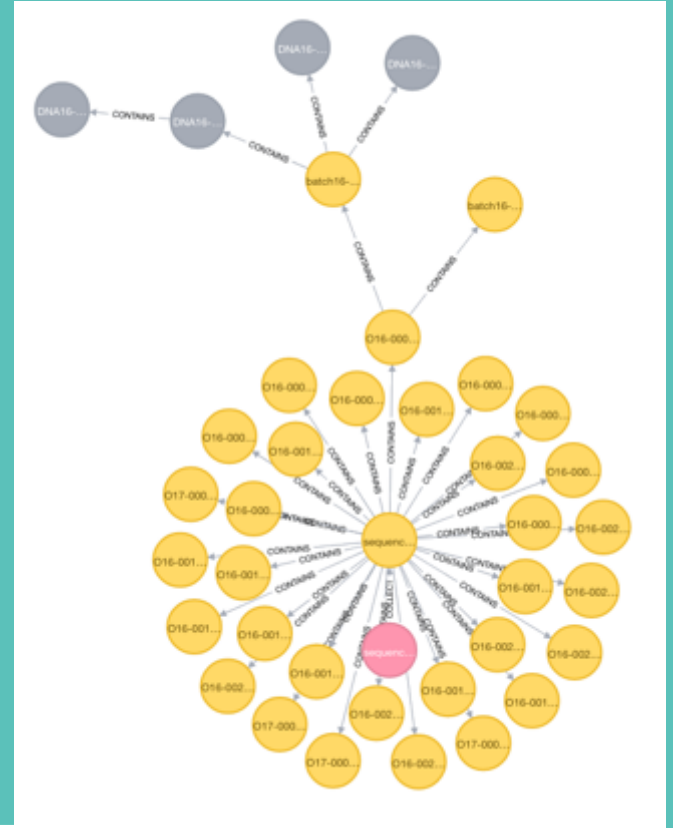
- run on a single instance
- run on multiple instances
- run against multiples OSEs
- be executed from cli
- available only from web
- ...

<https://bio.tools>

The screenshot shows the bio.tools website interface. At the top, there is a navigation bar with the bio.tools logo, a search bar containing "Search tool and data services registry", and a button indicating "10533 tools". There are also "Login" and "Register" buttons. Below the navigation bar, the main content area displays three tool cards:

- CoVaCS**: A fully automated, highly accurate system with a web based graphical interface for genotyping and variant annotation. Extensive tests on a gold standard benchmark data-set-the NA12878 Illumina platinum genome- confirm that call-sets based on our consensus strategy are completely in line with those attained by similar command line based approaches, and far more accurate than call-sets from any individual tool. Tags: DNA polymorphism, Web application.
- COSMIC**: Curates comprehensive information on somatic mutations in human cancer. Full scientific literature curations are available on 83 major cancer genes and 49 fusion gene pairs. Biomart allows more automated data mining and integration with other biological databases. Annotation of genomic features has become a significant focus. It integrates many diverse types of mutation information and is making much closer links with Ensembl and other data resources. Tags: Genetic variation, Oncology, Data mining, Database management, Database portal.
- CONTRAFold**: A novel secondary structure prediction method based on conditional log-linear models, a flexible class of probabilistic models which generalize upon SCFGs by using discriminative training and feature-rich scoring. The result closes the gap between probabilistic and thermodynamic models, demonstrating that statistical learning procedures provide an effective alternative to empirical measurement of thermodynamic parameters for RNA secondary structure prediction. Tags: Structure prediction, RNA, Physics, Command-line tool.

Metadata matters



Enabling computational reproducibility

Documented, portable, comparable, automated software.

Following this practices:

- develop code under revision control systems
- isolate execution by containers
- record metadata and data provenance
- automated IT steps

Your software should easily run on your own laptop, as well as on HPC cluster, as well as on commercial or research clouds.



git



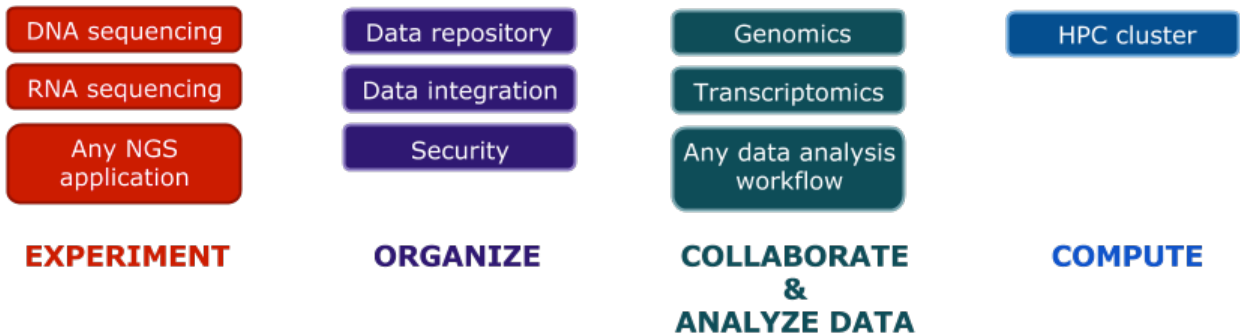
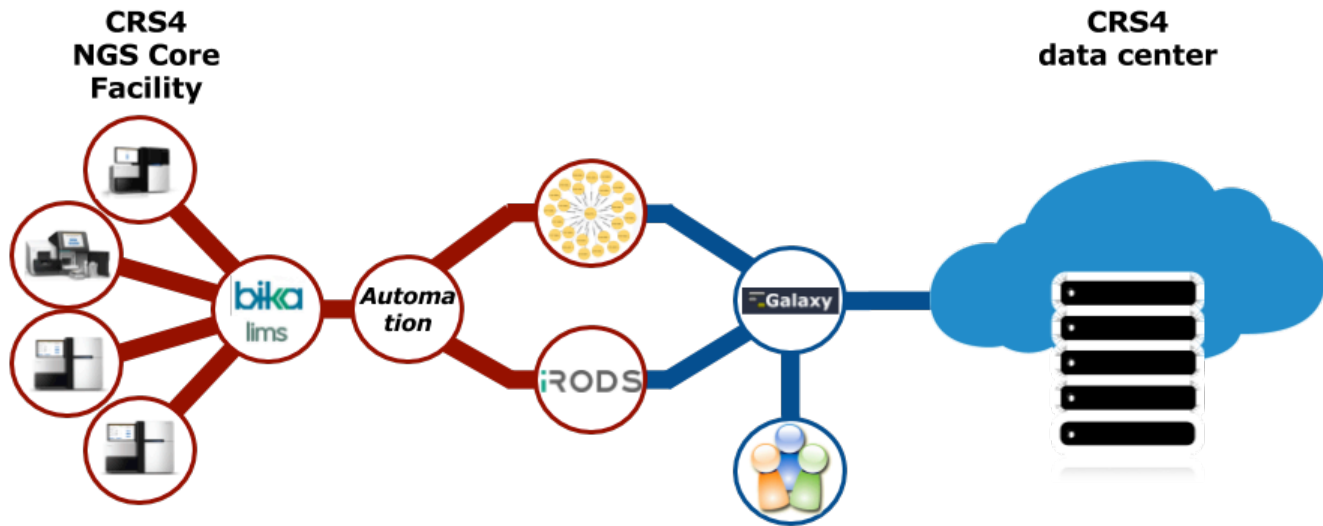
Biocomputing Infrastructure

@CRS4:

Data Management and

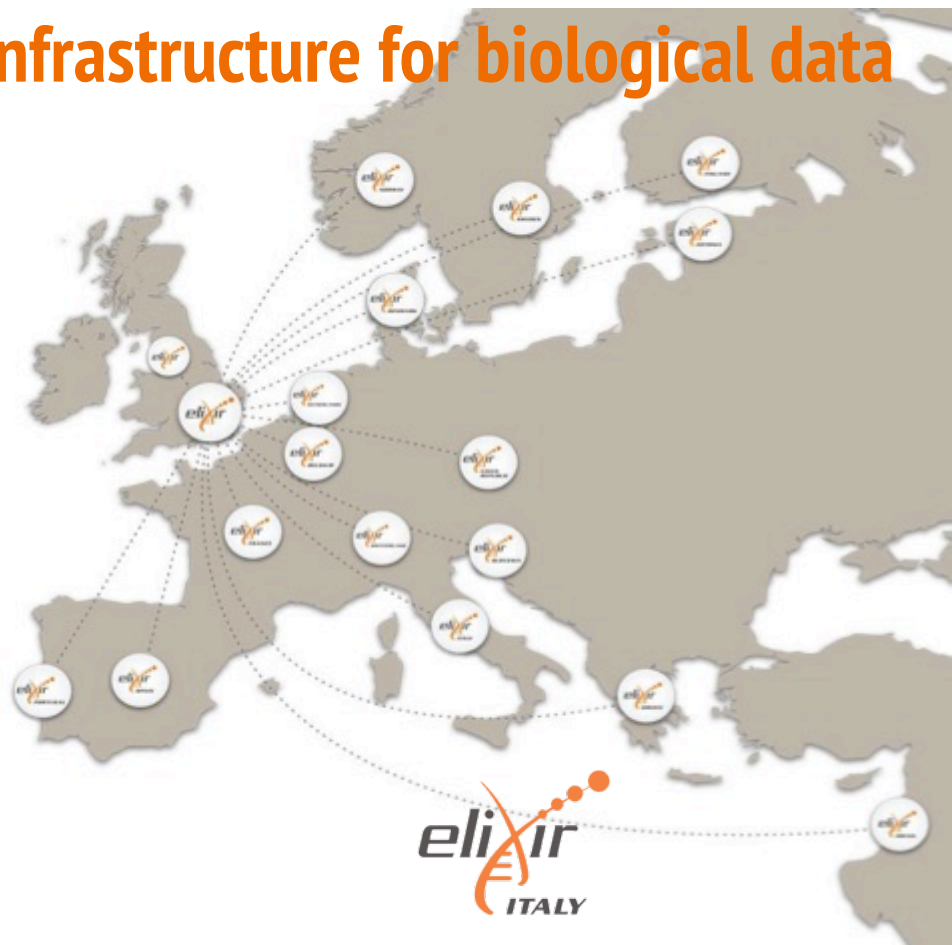
Computable Framework for NGS

studies



ELIXIR: the European Research Infrastructure for biological data

- ELIXIR connects national infrastructures and EMBL-EBI
- 17 Member states + EMBL
Belgium, Czech Republic, Estonia, Denmark, Finland, France, Israel, Italy, Netherlands, Norway, Portugal, Slovenia, Spain, Sweden, Switzerland, UK, Ireland
- ELIXIR deliver services through national ELIXIR Nodes
- ELIXIR Nodes build on national strengths and priorities



Training experience



Galaxy for Bioinformatics tool developers (ELIXIR IIB Training Programme)	Development	Cagliari	3-5 July 2017
Exome analysis using Galaxy (ELIXIR IIB Training Programme)	Exome analysis	Milano	19-20 September 2016
NGS data analysis with Galaxy	Exome-Seq, RNA-Seq, microbiology	Pula, CA	18-20 November 2014
NGS data analysis with Galaxy	Exome-Seq, RNA-Seq, microbiology	Cagliari	18-20 June 2014 23-25 September 2014
NGS data analysis with Galaxy	Galaxy introduction, QC, microbiology	Bologna	30-31 January 2014
NGS data analysis with Galaxy	Exome-Seq, RNA-Seq, microbiology	Pula, Ca	8-11 June 2013 18-21 October 2013
NGS data analysis with Galaxy	Bacterial sequencing (resequencing & <i>de novo</i>), metagenomics	Teramo	12-16 November 2012
Master in Bioinformatics	ChIP-Seq, RNA-Seq, systems biology, protein 3D modelling	Cagliari	6 September, 20 October 2012

100+ researchers from 20+ institutions attended our Galaxy courses.



Biosciences

Data Engineering & Computing

Focusing on:

- integration, interpretation and data intensive analysis of massive and heterogeneous biological data

Our work has the objective of enabling technologically advanced computational and experimental platforms for life sciences and clinical research applications.

